AIMS OF THE THESIS

- This work aims to use machine learning techniques for the classification of specific parts of web page content.
- Visual representation of web pages serves as input for training of machine learning models.
- The model architecture is based on graph neural networks.
- The advantage of the proposed and implemented approach is information extraction independent of the structure and language of a web page.

ABSTRACT

- The page is first converted to a graph model where each node contains a number of visual features.
- Leaf nodes can contain required information and are the target of classification.
- Target nodes with the context of a graph enter the Embedder, which consists of graph neural layers.
- From the Embedder comes an embedding that represents a node relative to the desired task and graph context.
- The embedding is classified by the Classifier, resulting in probabilities that indicate the degree to which the node contained information of interest.

MOTIVATION

- Identification and storing of important data from the website.
- Use of visual and spatial information.
- Quick retrieval of key information from any web page in a given category.
- Save time when searching for information.
- Possibility of further processing of automatically obtained information.

RESULTS

- The best trained models achieve an accuracy of 98.38% with an F1 score of 0.9837.
- The success rate of finding the desired information is up to 97.83%.
- Model was compared with other work using a predictive accuracy metric: